Accept or Decline? An Analytics-Based Decision Tool for Kidney Offer Evaluation

Dimitris Bertsimas, PhD, Jerry Kung, MAST, Nikolaos Trichakis, PhD, David Wojciechowski, DO, and Parsia A. Vagefi, MD

Background. When a deceased-donor kidney is offered to a waitlisted candidate, the decision to accept or decline the organ relies primarily upon a practitioner’s experience and intuition. Such decisions must achieve a delicate balance between estimating the immediate benefit of transplantation and the potential for future higher-quality offers. However, the current experience-based paradigm lacks scientific rigor and is subject to the inaccuracies that plague anecdotal decision-making. Methods. A data-driven analytics-based model was developed to predict whether a patient will receive an offer for a deceased-donor kidney at Kidney Donor Profile Index thresholds of 0.2, 0.4, and 0.6, and at timeframes of 3, 6, and 12 months. The model accounted for Organ Procurement Organization, blood group, wait time, DR antigens, and prior offer history to provide accurate and personalized predictions. Performance was evaluated on data sets spanning various lengths of time to understand the adaptability of the method. Results. Using United Network for Organ Sharing match-run data from March 2007 to June 2013, out-of-sample area under the receiver operating characteristic curve was approximately 0.87 for all Kidney Donor Profile Index thresholds and timeframes considered for the 10 most populous Organ Procurement Organizations. As more data becomes available, area under the receiver operating characteristic curve values increase and subsequently level off. Conclusions. The development of a data-driven analytics-based model may assist transplant practitioners and candidates during the complex decision of whether to accept or forgo a current kidney offer in anticipation of a future high-quality offer. The latter holds promise to facilitate timely transplantation and optimize the efficiency of allocation. (Transplantation 2017;101: 2898–2904)

The current demand for kidney transplantation continues to outpace the supply; since 2002, the number of candidates on the waitlist has nearly doubled, from just over 50 000 to more than 96 000 by 2013. In contrast, living donation rates have decreased since 2002, and the deceased donation rate has not increased in recent years. To further compound the problem of organ supply and recipient demand, there is an unacceptably high deceased donor organ discard rate—up to 50% for high kidney donor profile index (KDPI) organs and disproportionately high for increased infectious risk donors when compared with noninfectious risk donor counterparts. The decision to accept a deceased donor organ involves a complex series of well-intentioned “what if” calculations on the part of the accepting physician that rely on experience and intuition, instead of on data and patient outcomes.

The desire to provide patients with the highest quality organ holds potential to become our Achilles heel. The desire to maximize the outcome for an individual patient can increase discard rates and overlooks the aggregate benefit achieved with an overall expansion of organ transplantation. It is at this crossroad that the physician’s semiquantitative calculus and “gut” instinct come to the forefront and would benefit from a data-driven tool to assist in this complex decision-making algorithm.

We developed an analytics tool to assist physicians in deceased-donor kidney acceptance decisions using a model based on patient-specific data regarding previous organ offers. In doing so, we sought to estimate an individual patient’s waiting time, in the context of a current active organ offer, until the time to the next offer for a kidney of a KDPI threshold or lower. If widely adopted, such a tool could not only provide reassurance for organ offer acceptance but may also help improve the overall efficiency of kidney transplant allocation.

MATERIALS AND METHODS

We aimed to address the following prediction problem: what is the probability of a patient being offered a deceased-donor
kidney of some KDPI threshold (or lower) within some timeframe, given their individual characteristics? We considered different specifications with KDPI thresholds of 0.2, 0.4, and 0.6, and timeframes of 3, 6, and 12 months. Because KDPI is an imperfect quality metric, we only considered offers not previously declined by more than 50 candidates in all 3 specifications. Nonetheless, our methods can readily accommodate alternative KDPI thresholds, timeframes, and previous offer decline cutoffs. Individual patient characteristics relied on waitlist registry information and included Organ Procurement Organization (OPO) of listing, blood group, accumulated wait time, calculated panel-reactive antibody (cPRA), age, DR antigens, and information regarding past offers (date, KDPI, and match run sequence number from previous offer[s]).

Data

Waitlist and deceased donor information for the period April 1994 to September 2013 was obtained from the Organ Procurement and Transplantation Network (OPTN) Standard Transplant Analysis and Research data set. Match run information for deceased donor kidneys that were eventually accepted by a patient during the period March 5, 2007, to June 30, 2013, was obtained from the OPTN Potential Transplant Recipient (PTR) data set. For the period that our PTR data spans, we retrieved 38 806 match runs and their associated donor information; of the 38 806 match runs, 535 were explicitly documented as directed donations or assigned for multi-organ transplants and were removed. For the same period, we retrieved waitlist information for patients who were active for at least some fraction of it (N = 287 283).

Methods for Wait Time Prediction

The prediction problem was addressed using data analytics methods that are trained on historical data. Specifically, we used random forests, a well-studied method in the machine learning literature that has been widely applied to prediction problems. For each specification, our random forest models predict the probability with which a patient would be offered a kidney within the specified timeframe (the dependent variables), given their characteristics and offer histories (the independent variables). To be able to make such predictions, these models were first trained using historical observations of independent variables and their associated dependent variables. Once trained, the models then predicted as output the dependent variables, given (potentially previously unseen) observations of the independent variables.

Observations and Independent Variables

An observation corresponded to a patient registered at a transplant center in 1 of the 10 most populous OPOs and a random date within the period of our PTR data set. For each patient, 8 random dates were sampled—therefore generating 8 observations—for every year accrued on the waitlist.

For each generated observation, the dependent variable was computed as 1 (0) if the patient was (not) offered a kidney of specified KDPI within the specified timeframe from the observation date. Observations for which patients were inactive for more than 90% of the follow-up timeframe or became inactive because of accepting a kidney of KDPI higher than the specified threshold during the follow-up timeframe were censored. Consequently, the number of observations per patient was essentially proportional to the time they had been active on the waitlist.

The independent variables computed for each observation were:

1. The patient’s OPO;
2. The patient’s blood group;
3. The patient’s accumulated wait time on the waitlist;
4. The expected number of DR mismatches that the patient would have with the next kidney to be offered;
5. An indicator variable of whether the patient was pediatric;
6. An indicator variable of whether the patient was sensitized (cPRA > 0.8);
7–9. The number of kidneys of KDPI ≤ 0.4; 0.4 < KDPI ≤ 0.7; and 0.7 < KDPI ≤ 1.0 offered to the patient before the observation date with organ sequence number being at most 500, divided by the number of days the patient was active before the observation date;
10. The patient’s average organ sequence number in the match run for offers they received before the observation date;
11. The mean number of DR mismatches with kidneys offered to the patient before the observation date;
12. The average KDPI of kidneys offered to the patient before the observation date;
13. The patient’s current cPRA value at the observation date;
14. A factor variable indicating the month of the observation.

Each independent variable added predictive ability to the model. Because patients belonging to different OPOs and blood groups were subject to different rates of waitlist congestion and donation rates, variables 1 to 2 provided context for the remaining variables. Variables 3 to 6 contributed directly to the points system that is used to rank patients in match runs. To compute variable 4, the patient’s DR antigens were compared with that of all previous donor kidneys (from the same OPO) before the observation date, and then the average number of mismatches was calculated. This captured the points that the patient was expected to receive due to DR antigen matching. Variables 7 to 9 measured the frequency of receiving offers before the observation date, for each of the quality categories considered. Because it was observed that inclusion of all organ offers greatly reduced predictive power, only offers with sequence number at most 500 were considered; this threshold determination was made via validation. Additionally, as candidates are not offered kidneys during periods of inactivity, variables 7 to 9 self-accounted for the time the patient had been inactive. Variables 10 to 12 provided an indication of the patient’s ranking and of the quality of the organs offered at previous match runs. Variable 13 informed the model with an indication of how many procured kidneys would be compatible with the patient. Variable 14 captured fluctuations that were observed between the number of deceased donors in any given month.

It is worth noting that, except for the last variable controlling for month of the year, all the other variables were specific to the patient, thus allowing the model to make personalized predictions for waitlist candidates.

Performance Evaluation

Observations spanning the entire period of available match-run data were split into training, validation, and testing sets in chronological order, a standard performance evaluation procedure in the data analytics literature. For example, for the 12-month timeframe specification, the training set included all observations whose dates fell between May 1, 2007, and April 30, 2008. The validation set included observations dated between May 1, 2009, and April 30,
2010, whereas the testing set included observations dated between May 1, 2011, and April 30, 2012. There was a 12-month gap between each of these cut-off dates as the dependent variable captured information up to 12 months after the observation date. A random forest model was fit on the training set, and the out-of-sample area under the receiver operating characteristic curve (AUC) was computed for the validation set. AUC remains a standard metric for determining a model’s ability to distinguish between 2 outcomes in a binary prediction model. In this setting, for example, a model with an out-of-sample AUC value of 0.60 can correctly distinguish between a randomly selected negative example and a randomly selected positive example 60% of the time.

We varied the number of trees and the number of variables per tree as the parameters to the random forest. The parameter values eventually selected were the ones that yielded the highest AUC for the validation set. The model was finalized by retraining it on combined training and validation sets for the selected parameter values. To simulate real-world out-of-sample testing procedures, performance was only evaluated on the testing data set using the finalized model.

Accuracy Within Clusters of Similar Patients
To evaluate the accuracy of the model, out-of-sample predictions over clusters of “similar” patients were evaluated as follows. First, observations in the testing set were clustered by OPO, blood group, the number of years the patients had accrued on the waitlist and their DR antigens. Second, for each cluster and each dependent variable, we compared the variable’s average value over the cluster that our model predicted with the actual average value in the testing set. Put differently, this allowed the comparison of the average number of candidates that our model predicted to receive an offer of the specified KDPI within the specified timeframe with the average number of candidates that received such an offer in practice.

Creating Data Sets With Rolling Horizons to Assess Adaptability of the Method
To assess the accuracy of our models as the amount of data collected increases, we created prediction models on subsets of the data starting with just the first 3 years of data, gradually increasing to the entire data set.

Figure 1A illustrates how 3 years of data can be used for the 3-month timeframe specification. A 3-month buffer after each of the training, validation, and testing sets was used since predictions within 3 months were being investigated. When the amount of data available is increased to 4 years, the data can then be spliced into training/validation/testing in multiple ways (Figures 1B and C). For a given amount of data and a timeframe specification, all possible splits of training/validation/testing sets that are feasible were created, limiting both validation and testing sets to either 6 months, 9 months, or 1 year. For each possible split, prediction models were trained and their out-of-sample AUC values obtained.

Disclaimer
PTR data were obtained under IRB approval (F23797-101) from the Harvard University Committee on the Use of Human Subjects in Research. All identifiers were removed upon data receipt for the purposes of this study.

This study used data from the OPTN. The OPTN data system includes data on all donor, waitlisted candidates, and transplant recipients in the United States, submitted by the members of the OPTN, and has been described elsewhere. The Health Resources and Services Administration, US Department of Health and Human Services provides oversight to the activities of the OPTN contractor.

RESULTS

AUC
The model demonstrated excellent out-of-sample performance, with AUC values of 0.86, 0.88, and 0.87 when predicting the probability of receiving an offer with KDPI of 0.2 or less in 3, 6, and 12 months, respectively. Table 1 also reports out-of-sample AUC results for KDPI thresholds of 0.4 and 0.6 for all timeframes considered. We observed that the model’s performance remained consistent for all thresholds considered.

<table>
<thead>
<tr>
<th>KDPI threshold</th>
<th>3 Months</th>
<th>6 Months</th>
<th>12 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDPI ≤ 0.2</td>
<td>0.862</td>
<td>0.881</td>
<td>0.866</td>
</tr>
<tr>
<td>KDPI ≤ 0.4</td>
<td>0.878</td>
<td>0.881</td>
<td>0.870</td>
</tr>
<tr>
<td>KDPI ≤ 0.6</td>
<td>0.880</td>
<td>0.880</td>
<td>0.881</td>
</tr>
</tbody>
</table>
To illustrate which variables contributed the most to the random forest model’s predictive power, we present variable importance plots in Figure 2 for the KDPI of 0.4 or less specification. We observed that average organ sequence number remained the most important variable for each of the 3 timeframes. Years on the waitlist, intensity of previous kidney

**FIGURE 2.** Variable importance plots for 3-, 6-, and 12-month random forest prediction models.
Accuracy Within Clusters of Similar Patients

Figure 3 presents an accuracy comparison between actual and predicted fractions of patients for blood group O in OneLegacy, CA (OPO code: CAOP) receiving offers of KDPI of 0.4 or less within 6 months, broken down by clusters based on years accumulated on the waitlist. For example, for the cluster of patients who had been waiting 5 years, our model predicts a 2.6% fraction of them to receive a KDPI of 0.4 or less kidney offer within 6 months, whereas the actual fraction was 3.9%. A similar comparison for all the blood groups within CAOP is demonstrated in Figure 4.

Figure 5 demonstrates a similar comparison for patients in LiveOnNY, NY (OPO code: NYRT) clustered by both wait time and DR antigen profile. Patients with more compatible DR antigen profiles tended to have higher chances of receiving a KDPI of 0.4 or less offer, both in practice and in our model predictions, demonstrating that the models provided reasonable accuracy accounting for a variety of specific patient characteristics. For the example depicted in Figure 5, the most compatible 33% of patients (ie, patients with fewer than 1.61 expected DR mismatches) were predicted to be 4.9% more likely to receive a KDPI of 0.4 or less organ in year 6 compared to the least compatible 33% of patients (ie, patients with more than 1.71 expected DR mismatches), whereas the actual difference was approximately 5.4%.

Example of Intended Usage

A deceased donor kidney with KDPI of 0.55 is offered to patient J.S. who has accumulated 5 years and 4 months of waiting time. J.S.’s blood group is O, cPRA is 5%, and DR antigens are 7 and 13. Two months ago, J.S. was offered a kidney with KDPI of 0.83, and on that match run was sequence 153. What is the probability that J.S. will be offered a “high-quality” kidney (KDPI ≤ 0.4) within the next 6 months?

Table 2 demonstrates the models’ predictions, for several different OPOs, of J.S.’s probabilities of receiving a KDPI of 0.4 or less offer within the next 6 months. The actual probabilities that patients like J.S. received such offers in practice are also reported. For each OPO considered, the models’ predicted results were very close to the actual probability. For example, in PADV, 19.0% to 24.2% of patients like J.S. received KDPI of 0.4 or less offers within 6 months; the model’s prediction for J.S. was 22.7%.

On the Adaptability of the Prediction Method

The OPTN kidney allocation system was revised in December 2014. Due to the new rules instituted and the potential for even further modifications, it is reasonable to expect that new prediction models must be trained as data is accrued under any new allocation system to achieve high-quality out of sample results. However, the extent of data required to achieve high-quality AUC results remains to be determined. To further investigate this, a rolling horizon was created starting with 3 years of data and evolving to the complete data set range of 6 years, with observed changes in AUC under an increasing amount of available data. The construction of these rolling horizons is detailed in the methods section.

DISCUSSION

Transplant practitioners and waitlisted candidates are confronted with the crucial decisions of whether to accept or decline an offered organ, often in the setting of a limited time window to decide. The current paradigm relies on a practitioner’s experienced-based approach, which lacks scientific rigor and is subject to unreliable and irreproducible decision-making. We sought to develop an analytical tool...
that could not only assist practitioners in their organ acceptance decisions at these critical moments, but also serve as an educational tool for candidates awaiting kidney transplantation. In defining the future organ offer landscape in a patient-specific format, we hope to not only provide transplant practitioners the ability to achieve expedited, evidence-based, decision-making for organ selection but also to provide an interactive educational tool for transplant candidates to further their understanding of an additional aspect of the risk/benefit ratio associated with organ offer acceptance—specifically the factor of additional waiting time.

The problem of deciding whether to accept or decline an offered kidney, to maximize a patient’s overall outcome and maintain efficiency in the allocation system, can be thought of as an optimal stopping problem: how long should a patient wait before accepting an offered kidney that best suits their needs? By developing data-driven predictive analytics models that provide information about the probability of being offered high-quality organs in the future, we are closer to solving the optimal kidney acceptance problem. Indeed, the next extension of these predictions will rely on developing dynamic optimization models capable of assessing the accept/decline tradeoff by incorporating data regarding the patient’s current condition on dialysis. Ultimate application will require an initial assessment in a controlled prospective fashion to ascertain the utility and benefit of a decision support tool for kidney offer acceptance.

From a more global perspective, properly contextualizing the probability of receiving a future higher quality kidney offer, in comparison to a candidate’s current offer, will facilitate the organ offer decision-making process. Although this conceivably holds potential to allow for a more efficient process of donor and recipient matching and ultimately contribute to fewer discarded kidneys, this remains to be determined and will require application of the predictive model in a prospective fashion.

The decision to proceed with organ acceptance for transplantation incorporates decision-making based on a multitude of factors, including: donor quality, recipient condition, the incompatibility between the 2, as well as center-specific or practitioner-specific influences. Importantly, the calculus associated with accepting a deceased-donor organ offer revolves around the quality of the organ at hand compared to the anticipated organ quality of the next offer in the context of a patient’s waiting time and comorbidities. However, the various weight applied to these individual factors likely varies based on which organ is being transplanted. Volk et al, recently demonstrated an organ support tool to assist practitioners with organ offers for candidates waitlisted for liver transplantation with the hopes of improving patient survival in liver transplantation. The decision to proceed with a kidney offer likely differs from that of other solid organ transplants (ie, liver, heart, or lung), for which a more immediate “life or death” situation guides the decision-making process. Indeed, for these nonrenal organs the immediacy of the candidate’s demise while awaiting the next “higher quality” offer likely holds stronger influence in the decision-making process for offer acceptance when compared with that for kidney transplant candidates who can maintain stability (although not indefinitely) through renal replacement therapy. Given the kidney transplant candidates’ option for dialysis, and reduced risk of demise precluding transplant while waiting for the “next” offer, we hypothesize that a decision support tool may have its greatest impact in renal transplantation.

Our data set included deceased donor kidneys between the years of 2007 and 2013. On December 4, 2014, OPTN introduced a new kidney allocation system. Although the data presented herein, and thus the results, were based on the prior

<table>
<thead>
<tr>
<th>OPD</th>
<th>Actual probability</th>
<th>Model’s prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>California (CAOP)</td>
<td>3.4-3.7%</td>
<td>3.8%</td>
</tr>
<tr>
<td>Illinois (ILIP)</td>
<td>8.3-15.7%</td>
<td>11.7%</td>
</tr>
<tr>
<td>New York (NYRT)</td>
<td>4.2-8.6%</td>
<td>7.9%</td>
</tr>
<tr>
<td>Pennsylvania (PADV)</td>
<td>19.0-24.2%</td>
<td>22.7%</td>
</tr>
<tr>
<td>Texas (TXGC)</td>
<td>23.3-26.2%</td>
<td>30.7%</td>
</tr>
</tbody>
</table>

CAOP, OneLegacy, CA; ILIP, Gift of Hope Organ & Tissue Donor Network, IL; NYRT, LiveOnNY, NY; PADV, Gift of Life Donor Program, PA; TXGC, LifeGift Organ Donation Center, TX.
allocation system, the new system introduced resulted in a more uniform allocation process across OPOs, diminishing the effect of regional variances in allocation. Our models are already tailored to account for OPO-specific differences that were observed previously; thus, decreasing the variation among OPOs would likely strengthen the predictive power of our models. Despite these OPO variances within our data set, the current results show promise for extending these methods for predictions under the new allocation rules once enough data has been accrued. Indeed, our results on the adaptability of the prediction model demonstrate that under the new allocation system, sufficient data has yet to be collected to implement this prediction method at 3, 6, and 12 months. It is anticipated that the combination of further accrual of data under the new system, with a reduction in variability between OPOs, may allow for improved AUC results compared to our current findings.

It is important to note that our current model is proof of principle of the power for an analytics decision tool to enhance organ offer decision making. It is thus anticipated that with time, the data sets generated from the new allocation schemes will be applied in the same context. Indeed, the current models need to be trained on historical data only once; thereafter, the process of making predictions for specific patients can be computed almost instantaneously.

The data set examined consisted only of kidneys that were eventually accepted by a candidate, and thus did not allow for examination of kidneys that were discarded; the latter leads to a degree of biasing in our data set. Incorporation of discard data will allow better predictions about the probability with which patients might be offered lower-quality kidneys, and would thus strengthen the overall prediction capabilities of the model. It should also be noted that although the resultant models using random forests are not easily interpretable, the model. It should also be noted that although the resultant models using random forests are not easily interpretable, the model would thus strengthen the overall prediction capabilities of the model. It should also be noted that although the resultant models using random forests are not easily interpretable, the model would thus strengthen the overall prediction capabilities of the model.

We observe that random forest significantly outperforms classification trees using logistic regression as our prediction model and found its AUC to be weaker than the AUC of the random forest model, see Table 4A. Moreover, we also found that the logistic regression was significantly less accurate for individual predictions, see Table 4B.

Application of methods from the data analytics and machine learning literature has allowed for the development of predictive models to provide personalized probabilities of receiving kidneys of specified quality within timeframes of 3, 6, and 12 months. We have demonstrated that these models have strong predictive ability out-of-sample, and thus can be used in real-world scenarios. Importantly, our models are strictly data-driven and rely on minimal assumptions. By informing physicians and transplant candidates with these probabilities, we hope to progress from intuition-based decision-making, to actionable insights gleaned from data.

ACKNOWLEDGMENTS
The authors thank Phebe Vayanos for her help with compiling the relevant data sets.

The data reported here have been supplied by UNOS as the contractor for the OPTN. The interpretation and reporting of these data are the responsibility of the author(s) and in no way, should be an official policy of or interpretation by the OPTN or the U.S. Government.

REFERENCES